DOCUMENT RESUME

ED 072 561 EA 004 936

AUTHOR	Dupuis, Mary M., Ed.; Mitzel, Harold E., Ed.
TITLE	The Role of Evaluation and Assessment Within the National Institute of Education. Report of a Planning Conference for the NIE Planning Unit (Washington, D.C., January, 1972).
INSTITUTION	National Inst. of Education (DHEW), Washington, D.C. Planning Unit.
REPORT NO	NIE-C107
BUREAU NO	BR-1-7059
PUB DATE	Jan 72
GRANT	OEG-0-71-3636(515)
NOTE	8p.
EDRS PRICE	MF-$0.65 HC-$3.29
DESCRIPTORS	Agency Role; *Conference Reports; *Education; *Educational Planning; Educational Problems; Educational Research; *Evaluation; Evaluation Needs; *Government Role; Research Methodology
IDENTIFIERS	Assessment; *National Institute of Education; NIE

ABSTRACT

This publication reports conference discussions covering some specific problems that confront education, several general and specific problems within research and evaluation, and several organizational and functional characteristics of the research community and the National Institute of Education that may be expected to affect the quality of their interaction. Specific educational problems discussed include those concerning disadvantaged students and maintenance of the quality of education for all students. Some more specific educational issues such as what evaluation should measure are also discussed. A related document is EA 004 935. (Author/DN)

ED 072561

U S. DEPARTMENT OF HEALTH,
EDUCATION & WELFARE
OFFICE OF EDUCATION
THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL OFFICE OF EDUCATION POSITION OR POLICY

Report No. C107

THE ROLE OF EVALUATION AND ASSESSMENT WITHIN THE NATIONAL INSTITUTE OF EDUCATION

Report of a Planning Conference
for the NIE Planning Unit

January 1972

Washington, D. C.

Thomas S. Barrows, Educational Testing Service
C. W. Churchman, University of California, Berkeley
Robert B. Davis, Syracuse University
Edmund W. Gordon, Teachers College, Columbia University
Philip W. Jackson, University of Chicago
Dean Jamison, Stanford University
John M. Mays, Office of Science and Technology
Samuel Messick (Chairman), Educational Testing Service
Eleanor H. B. Sheldon, Social Science Research Council
Harry Silberman, NIE Planning Unit
Richard E. Snow, Stanford University
Joachim Wohlwill, Pennsylvania State University

NATIONAL INSTITUTE OF EDUCATION
Planning Unit

EA 004 936

This series is prepared under Project No. 1-7059, Grant No. OEG-0-71-3636(515), for the U.S. Office of Education's National Institute of Education Planning Unit, Dr. Harry Silberman, Director.

Edited by Mary M. Dupuis, Ph.D.; Harold E. Mitzel, Ph.D., Project Director, College of Education, The Pennsylvania State University.

This Planning document was submitted to the NIE Planning Unit. Views or conclusions contained in this study should not be interpreted as representing the official policy of the NIE Planning Unit, Office of Education, United States Department of Health, Education and Welfare.

THE ROLE OF EVALUATION AND ASSESSMENT WITHIN THE NATIONAL INSTITUTE OF EDUCATION

INTRODUCTION

In January, 1972, a two-day seminar was held in Washington, D.C., to discuss issues of importance to the formation and conduct of the National Institute of Education (NIE) and to the field of educational research and evaluation. Discussion ranged widely, covering some specific problems that confront current education, several general and specific problems within research and evaluation, and several organizational and functional characteristics of the research community and NIE that may be expected to affect the quality of their interaction.

Education, research and evaluation were all broadly conceived. The discussants repeatedly stressed that NIE must consciously extend its purview beyond institutional programs of education to include other situations in the general environment where learning and psychological growth are or potentially could be salient outcomes. Similarly, research should include methods and sources of knowledge that are currently viewed by some as "ascientific" according to the rigorous research paradigms and standards for data borrowed from research in the physical sciences. Evaluation must also be extended beyond the measurement of processes and their effects to the attribution of value to treatments or programs. This latter point received especially heavy emphasis. Since multiple and competing value structures occur in a pluralistic society, competing criteria of effectiveness arise, and choices among them involve the acceptance or rejection of values, whether that issue is addressed directly or not. Direct consideration of social values was favored throughout the evaluation process.

EDUCATIONAL PROBLEMS

One major problem that currently confronts education–perhaps the most compelling one–is the existence of what is termed a *disadvantaged* sector within our society. Although it is difficult to find viewpoints that dispute the magnitude of the problem, it is equally difficult to find adequate educational or psychological conceptualizations of the problem and its etiology. It is also difficult to find–and perhaps this follows necessarily–a body of knowledge about specific symptoms of "disadvantagement" to aid in establishing priorities among proposed programs of social action.

Two distinctly different ways of viewing the problem seem to underlie the proliferation of terminology encountered in the educational and psychological literature (i.e., "disadvantaged," "deficit," "deprivation," "difference," etc.). A *deficit model* suggests that one scheme of valued characteristics should be applied in "judging" all individuals or groups. Those persons or groups possessing little of a desired characteristic are seen as deficient, and the suggested remediation usually is an attempt to supply what is wanting. On the other hand, a *difference model* holds that different persons and groups do and should have different strengths and weaknesses. Different schemes of valued characteristics are seen, then, as appropriate for different groups, and differences between groups or individuals may or may not be cause for remedial efforts. Furthermore, efforts at remediation may be chosen to bring about even greater differences between groups or differentiation within them. Such courses of action are built upon accepting diverse valued characteristics and nurturing this diversity.

An excellent comparison of the two models occurs if one considers how to approach the relatively poor educational performance of several groups that speak English as a second language. According to the deficit approach one would try to improve performance on traditional educational measures–especially of English language skills where the deficits are great. According to the difference model, one would more likely take an approach that is becoming more popular today. The difference in English language skills is noted, but priority is given to supporting bilingual skills through a program designed to strengthen both native and acquired languages.

The highly visible problem of the disadvantaged sector leads directly to consideration of the *quality of education* for all groups within the current society. A common approach to quality is to adopt a production model whereby inputs, processes, and outputs are related for various groups. Several limitations of this approach may be seen. First, it commonly adopts the deficit approach in assuming a common set of goals, dimensions of performance, or human characteristics to be appropriate across groups. Second, theories of development stressing psychological reorganizations do not fit within it well. These "stage" theories focus upon change through repeated differentiation and integration

of characteristics resulting in new structures and the emergence of seemingly new skills. Quality can be conceived of as variation in structure (both number of dimensions and configuration) or in distributions (central tendency, dispersion, and shape). The production model adopts much too narrow a view in considering only levels of performance (distributions).

As we mentioned above, not only is there uncertainty about what model of disadvantagement to emphasize and how to conceptualize its mechanism, but the construct itself is poorly defined. That is, there seems to be little agreement about what constitutes disadvantagement in the aggregate and about what set of characteristics should serve as a sign or indicator of its presence. For example, the most widely used indicator is probably socioeconomic status (SES), a conglomerate index whose constituent parts are quite variable across applications. It has been use; to establish eligibility for most major compensatory education programs, although a measure of educational deficiency or need seems intuitively much more attractive. To our knowledge, the rationale for using SES in this case is unstated and untested critically.

There are, of course, a great many studies concerned with interrelationships between measures of SES, race, educational attainment, and a host of other economic, social, and psychological characteristics. In general, however, the studies employ too few of the potential variables, and too little accrues across studies to give convergent (and discriminant) meaning to disadvantagement as a construct.

In a situation where so little is known of a major social problem, a primary focus and large-scale effort on the part of NIE seems necessary. First, NIE should exercise leadership in conceptualizing the problem. This effort should foster continued consideration of alternative conceptualizations of disadvantagement and quality, of the underpinnings of those conceptualizations in differing schemes of social values, and of the social consequences of basing action programs upon them.

A major effort must similarly be focused on understanding the structure of environmental impacts on human development. To move towards this we must ask what kinds and levels of stimuli are effective. Although we currently measure effects with a good deal of sophistication, our measures of environment are crude. Attempts must be made to increase both the number of environmental dimensions we can tap and the sensitivity with which we tap them.

RESEARCH ISSUES—VALUES AND PROBLEMS

It was mentioned initially that educational research must be broadly conceived. Just as conceptualizations of disadvantagement and quality education grow from systems of values and interests, so too do systems of learning about and knowing about education. The resulting variety of methodologies should all be considered to be under the umbrella of educational research. For example, lawyers, historians, and social philosophers develop and use bodies of knowledge about education that are strikingly dissimilar to the bodies of knowledge pursued by the quantitatively oriented, scientific researcher. Although strikingly different and less "scientific," and in some instances obviously value-laden, it would be difficult to show that these bodies of knowledge are less worthy of pursuit. The wisest course does not seem to lie with attempts to free one tradition and methodology from its underpinnings in value consideration or to pursue one discipline at the expense of others. Rather than a "pure" or value-free social science, one which triangulates social problems from a number of openly value-centered positions is preferred.

There is danger that any rigidly disciplined system of knowledge may become somewhat paranoid. That is to say, the discipline may deny knowledge derived by methodologies other than its own—especially when the offending conceptualizations and models disagree with those it has itself derived. In some cases this occurs in the field of educational research and is also true of areas within it.

Consider, for example, the ongoing argument concerning the evaluation of educational or instructional programs. This might be called the Evaluation-as-Evaluation vs. the Evaluation-as-Research argument. Disagreement seems to center on the degree to which procedures for evaluative studies should be prescribed, with further (perhaps resulting) disagreements on what these procedures should be. One school, as an instance, suggests that specific behavioral objectives should be stated in advance and that terminal performance should be judged in terms of the objectives. The other school insists that unintended outcomes and

side effects must be considered as well and that a program must be judged in terms of all the effects it produces, not just in terms of its intended effects. Intended effects are seen as necessary but clearly not sufficient to deciding whether to continue the treatment. More important, the evaluation-as-research school argues that we must go beyond an assessment of the size of effects to an examination of the processes that produce the effects if we are to understand educational treatments in sufficient depth and generality to apply them adaptively under adverse circumstances. The differences drawn between these two schools tend to oversimplify many distinct opinions; for illustrative purposes, however, it seems clear that the heuristic considerations of simplicity and transportability that recommend the evaluation-as-evaluation viewpoint may well encourage subsequent paranoia or apologies for a prescribed program's frequent inadequacy. Side effects are frequently found in "second-round" studies as are a host of other effects and moderators. They should be expected. If practical considerations do not allow designs that will unearth them initially, it seems wiser to explain the limitations to be expected from partial application of the research model than it is to apply the simpler evaluation-as-evaluation model. The immediate danger in the latter, of course, is that attention is not drawn to what is now known, and so decisions may be premature even if not incorrect. In addition, little knowledge about the effectiveness of general characteristics of treatments is likely to accrue unless some overlap of criterion variables is planned from study to study.

Beyond those issues suggested above, the question of educational research's impact (or lack of it) on the practice of education bears directly on the argument for broadening research. From the point of view of impact, educational research and evaluation may be viewed as functions rather than as disciplines or collections of disciplines. These functions are, of course, to provide knowledge that will lead to improved educational practice. Once again it becomes clear that a large number of disciplines may be relevant. An appraisal of the impact of previous research generally leads, however, to the conclusion that direct impact has been slight. It was, in fact, suggested as early as 1932 that research had inter-correlated all the variables of interest to no avail and that it therefore remained for educational philosophy and rhetoric to mold education in accordance with considerations of values and ethics. Research was viewed as a trivial technology and its trajectory seemed to go nowhere.

A more optimistic appraisal suggests that the impact of educational and psychological research upon education has been massive but indirect. Research leads to changed conceptions of human development and of the nature of the human being as a learner, which in turn leads to modifications in value perspectives and in educational goals and processes derived from them. These changed conceptions and values are reflected in the rhetoric of the agents of educational change and so in the changes themselves, albeit indirectly. For example, this viewpoint suggests that recent efforts in early childhood education have their roots in the research work of Piaget and others who have changed society's conceptions of early development and its potentialities. Similarly, Skinner's work in the conditioning of behavior changed conceptions of learning reflected in programmed instruction and in regimens of behavior modification. Skinner's interpretations of his own work, which attempt to introduce a new controlled basis for individual freedom, constitute an especially interesting attempt to hasten the indirect transmission of research findings in reshaping conceptions and associated values. As might be expected, critics of this attempt focus upon Skinner's interpretive processes rather than his research findings per se.

If one accepts the view that impact of research has been non-trivial though indirect, the problem of increasing its impact is recast. Previous attempts of "putting research findings into practice" have stressed a "dissemination model," rivaling a pipeline in simplicity. Research findings are thus made available to practitioners directly after minor efforts of aggregation and synthesis. While this is intuitively attractive in its simplicity, it doesn't seem to work. If we believe that the influence of research is indirect, we must seek to understand more about the process–that is to learn how the findings of the many disciplines functioning in educational research lead to new conceptions and values and thus change educational practice. A strong effort in this direction is suggested for NIE.

Finally, with regard to educational research generally, one point is common to considerations of values, functional and disciplinary orientations, and the influence of impact on practice. That is, the system in which all these phenomena inhere and interact is marked by continuous change. Because of this, the efforts suggested above should be expected to be continuing ones in order to achieve necessary understanding of each "new" educational system and its conceptual and empirical basis.

TRADITIONAL EVALUATION–ISSUES

Measurement data have been used throughout the educational enterprise in numerous ways that are quite distinct from and supplementary to their research uses. Traditional uses of measurement information include diagnostic and prescriptive decisions regarding individuals and groups. The use of similar information in self-evaluation, leading to choices between alternative courses of action by individuals, is becoming increasingly widespread. The term "guidance," applied to the selection between alternatives for individuals, focuses on the role of others in mediating the information or otherwise influencing the personal decision. Informed self-evaluation and individual choice, on the other hand, require much of the same assessment information as well as the development of skills in rational decision-making.

Evaluations of educational programs in most cases make use of measurement information. "Summative" evaluation provides an overall appraisal leading to statements about program effects or program value, while "formative" evaluation provides statements concerning the need to modify components of programs. This distinction between summative and formative seems an uncomfortable one when thus applied to types of evaluative activities or information. However, the suggestion that "summative" and "formative" distinguish two distinct types of decisions regarding programs, and, hence, two different *roles* (rather than forms) of evaluation, seems more useful and is in keeping with the fact that various types of information may contribute to decisions about programs.

Measurement information is increasingly used in continuous evaluation of educational systems. The purpose is to provide information about changes in the system and warning of the need for adaptive action. A broad array of information must be gathered, since the functioning of each important system component must be monitored. The aggregation of information at various levels and the ability to interrelate diverse elements of information are necessary to portray adequately the complex interactive functioning of the numerous components and contexts of educational systems.

Although these uses of evaluative information are traditional, they are not without problems. As noted above, different conceptions of guidance prompt the question of who should decide which course of action a person is to follow. Once a program is selected, questions of when to terminate the educational treatment are often handled in arbitrary ways. For example, when should absolute mastery be a standard for termination? Why are certification standards what they are? Why shouldn't individuals decide when to terminate or enter programs? These questions and large numbers of others are not answered by increasing the precision of measures, and their importance is not diminished by the scientific aura created by good data. The value-based assumptions leading to the establishment and acceptance of these standards should be continuously examined, and alternatives should be systematically considered.

A good example of the fruits of such a questioning approach may be furnished by recent considerations of the adequacy of so-called diagnostic testing. It has become increasingly evident that diagnoses that are not sensibly wed to prescriptions are of little if any clinical value. The central requirement of demonstrated trait-by-treatment interactions has recently received considerable attention, and we may consequently hope to progress more productively–or at least less wastefully and harmfully–in the future. Similar consideration of the rationale and value bases of traditional evaluative activities is suggested as an area for continuing major effort.

METHODOLOGY–SOME SPECIFICS

Within the context of broad educational issues, several narrower ones were also discussed at the conference; these are recommended for NIE's attention. The general rubric of *what to measure* applies to one set of them.

1. Measurement and evaluation usually focuses initially either upon discrete behaviors or upon psychological constructs predicated on previously observed consistencies and inconsistencies in numerous behaviors. Within evaluation, the behavioral objectives approach, and to some extent criterion-referenced measurement, emphasize discrete behaviors as the starting point and changes in those behaviors as the objective. The construct approach, on the other hand, emphasizes consistencies in behavior as the starting point and the acquisition of higher-order heuristics, such as rules, principles or strategies, as the objective (the latter

typically being viewed as constructs to explain or account for the behavioral consistencies). Differing basic conceptions of education may account for some of this apparent disagreement. If one views the purpose of education to be the teaching of specific behaviors, then the production and change of specific behaviors should be the educator's aim. If, on the other hand, education is seen as the development and modification of broader controlling mechanisms and principles, then these mechanisms should be of prime interest. The difference corresponds to the distinctions often made between education and training.

In practice, of course, the adoption of one point of departure does not determine the entire course of research or development. The developer or user of criterion-referenced instruments is not relieved of the responsibility of addressing at some point questions of generalizability and construct validity. Once again, NIE's stance should be to accept either approach and attempt to achieve balance within and across specific projects.

2. Measures of process and context are scarce and probably less adequate than measures of so-called educational output. Context and process measures are critical to research because one is often forced to study naturally occurring variation in educational systems. If it were possible to manipulate educational processes in the laboratory in a way that truly reflected complex system functioning, then the ability to monitor processes and to study contextual constraints on generalizability would not be as critical to achieving some understanding. Measures of this sort are also necessary for the system-monitoring mentioned above. The development of process and context measures is therefore suggested as having high priority for NIE.

3. Measures of developmental structures and of progression through stages or qualitative changes in these structures, are similarly scarce, and often are of unknown technical quality. Measures of discontinuous traits, of structures and restructured phenomena are suggested for further effort.

A second set of issues concerns *how to measure* and how to incorporate information (from measures or other sources) into decisions (evaluative or otherwise).

1. For the most part, measures of individual characteristics have focused on psychological traits conceived of as stable over time or situations. This has led to the application of temporal stability of scores as one criterion of the technical adequacy of measures. There are, however, psychological phenomena that are not conceived of as stable; in these cases, to apply the stability criterion would be totally wrong. A measure of mood, conceived as a transitory state, is a simple example in which temporal fluctuation would be desirable.

In addition to stability, conceptions of growth and decline are employed in measurement processes, and unless they are consciously compared with conceptions of the traits and processes thought to be operating, serious errors can occur. Measures of personality, affect, and stage phenomena are areas in which the warning should be clear. Methodologies developed to measure cognitive characteristics may not be directly applicable in other areas. A psychometrics developed for individual discrimination may be inappropriate to measurement of common mastery or stage attainment.

2. The current reliance upon normative interpretations of measures is not in keeping with the availability of scaling methods. Normative data may be adequate for a number of traditional educational uses, but a more thorough application of scaling methodology should permit substantive (as opposed to normative) comparisons of greater potential value in research and evaluation.

3. The scoring of constructed response should be made more efficient through concerted efforts to apply newly devised technology. The limitations of multiple-choice responses are well documented. In cases where that response format requires additional assumptions regarding traits or processes, the effort of constructed response scoring should be considered in the light of developing technology.

4. When we compare educational programs, a complex interaction occurs between the scales of criterion variables, the scales of value associated with criterion performance levels, and the index of similarity used. For example, in a comparison of a spelling program and a reading program the scales of reading and spelling measures, the scales of value of reading and spelling achievement, and the index of similarity used (difference in means or medians, comparisons of ranks, etc.) have individual and collective influence on whether we find the programs equally valuable. Further inquiry into the nature and effect of these interactions is

necessary if studies of comparative values of programs are to be supported.

THE ROLE OF NIE

The failure of research and development programs within the Office of Education (OE) is seen by many as the main reason for trying a "fresh start" with NIE. If the OE program is viewed as a failure, there is a lesson for NIE. The OE program satisfied neither researchers (who decried the lack of rigor) nor practitioners (who decried the lack of practical and immediately applicable results). This is but one example of the pluralism of values that we focused on initially and of the resulting multiple and conflicting demands inevitable within education.

As mentioned before, there is no single solution. The situation is not to be viewed as a problem to be solved once and for all but as a fact of life which NIE must cope with continually. The suggestion that the National Institutes of Health provide a viable model—and so a solution—is probably an unjustifiable one. The biomedical professions have standards of research excellence which are widely shared throughout that community. This is clearly not true for education and the research areas that undergird it—anthropology, sociology, economics, psychology, to mention only a few. The effect of creating NIE will be to further centralize educational research, a fact which may be expected to aggravate the situation by blunting the pluralism of the research community rather than capitalizing upon it and extending it. This may occur because communications with local groups holding local values will be artificially closed off.

The challenge is clearly great. A stance of openly accepting competing values and searching for solutions that can satisfy multiple and ever-competing demands is suggested as the wisest course. This, coupled with the role of intellectual leadership required to come to understand the diverse values that underlie an educational enterprise so diversely conceived, seems almost too much to ask. But still a further requirement must be made.

OE's current attempts to direct research and development activities limit the quality of these activities unwisely. In many instances, the research and development community is relegated to the role of purveyors of services centrally specified. Hence, their potential contributions to the initial conceptualization of problems are lost. That situation is not only inefficient; it is also distasteful to the research and development community. Perhaps the exercise of central leadership in the targeting of critical areas, as NIMH does, rather than attempts to direct specific efforts, would work well for education too by allowing both governmental priority setting and the fullest use of researchers' diverse skills. Clearly, NIE must have the sensitivity to "read" the subtle statements that convey a society's values, the tireless intellectual capacity to search out and consider alternatives satisfying competing demands, and the leadership to orchestrate the research and educational communities so that they contribute to their fullest capability.

Finally, the need to evaluate the efforts of NIE must be recognized from the outset. The magnitude of the challenges it faces and the consequences of even partial failure strongly recommend formative evaluation of NIE itself in the service of adaptive action. Several important issues must be addressed immediately so that these evaluative efforts may start as NIE begins to function. For example, the choice of internal versus external evaluation must be resolved (preferably through a combination of both), objectives and criteria must be sharpened, and the decision-relevance of information must be considered. It is certain that continuous adaptive action will be needed if the promise of NIE is to be fulfilled.